# 2.0 Data and Information Documentation for EMPACT Projects

All EMPACT projects are required to document their activities. This section provides guidance on the levels of documentation available to projects and recommends the types of information that are necessary to adequately describe EMPACT projects and data. Documentation is important because it helps users make informed use of data, provides consistency and project “memory” over time, and allows data to be shared and used in a variety of computing environments.

## 2.1 Levels of Documentation

EMPACT project must develop the following four kinds of documentation:

- *Project documentation*: documenting the highest level of information about the project;
- *Data set documentation*: clear information about what data is collected and how it may be accessed and used;
- *Data element documentation*: full definitions and specifications for each element collected and maintained in a data set; and
- *Database system documentation*: schematics and detailed information about how the data is managed.

Figure 2.1 on the next page provides a schematic diagram using typical database terminology to represent the relationship between data elements, data sets and a database.

### *2.1.1 Project Documentation*

Documentation at the project level should describe what the project is trying to accomplish, the process being used, and the resources required to undertake the project. This information will guide others who are developing similar projects and will assist those who wish to use the data to understand the methodologies used and interpret the meaning of the data.

### *2.1.2 Data Set Documentation*

Documentation at the data set level includes providing the source of the data, contact information for obtaining the data, and methodologies used in the creation of the data.

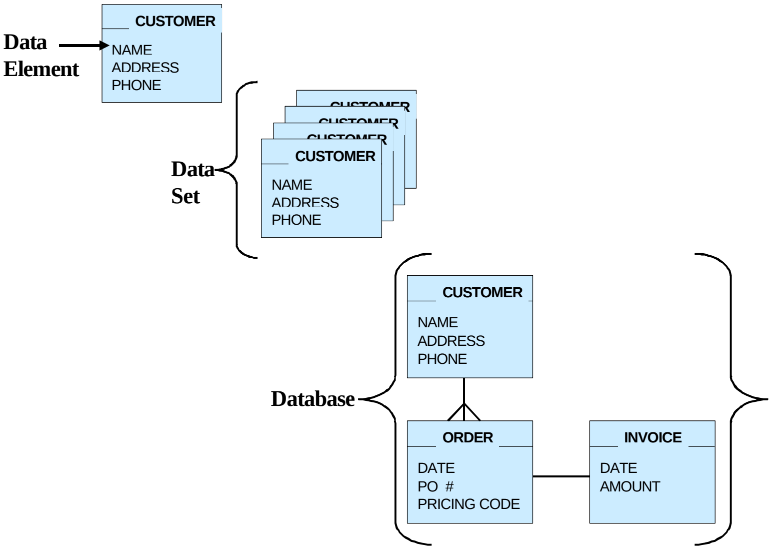

**FIGURE 2-1:** Representation of Data Documentation Terminology

Documentation of the data sets is intended to answer the following questions:

- What data were collected?
- Where were the data collected?
- When were the data collected?
- What is the data set named?
- Who were the data collectors?
- Who is the point of contact?
- What is the data set about?

Data set documentation, often referred to as metadata, is a collection of information about data. Metadata document the source(s), content, quality, lineage, structure, availability, and other important characteristics of a data set in the same way that a map legend describes a map.

The map legend contains information about the publisher of the map, the publication date, the type of map, a description of the map, spatial references, and the map's scale and accuracy. Metadata provide similar information about data sets. Metadata also describe the intent and potential uses of the data and assist in making decisions about the appropriate use of data. They document the data holdings of an agency to retain institutional knowledge through changes of personnel.

Metadata provide the information needed to allow users to make educated secondary use of project level information and to compare data across projects. When examining data from different sources, it is important to know when and where each data set was collected and whether similar methodologies were used. These factors have significant impact on any analysis derived from the data.

To gain an appreciation for the importance of metadata, consider a hypothetical user who wants to compare air quality data presented on the Web sites of two EMPACT projects monitoring air quality. This hypothetical user would want to know whether the information presented on a web page describing air quality in San Francisco can be interpreted the same way as similar information on a Web site related to Boston. In order to ascertain that the data have the same meaning, he would need to understand the methodologies used to collect data at the two locations to see if they are comparable. He would want to know the units of measure to ensure that he is comparing apples to apples. He would want to know when the data were collected, so that this information could be factored into his analysis. All of this information should be available in the metadata describing the data sets used.

To increase the value of metadata and allow a broad audience to read and interpret this information, standards have been developed for the creation of metadata. One such standard is the Federal Content Standards for Digital Geospatial Metadata and is described in Appendix C. EMPACT data set documentation follows the general approach taken by the Federal Geographic Data Committee.

### *2.1.3 Data Element Documentation*

A data element is a single unit of data that is the most basic level of documentation for a project. It cannot be divided into more fundamental segments of data and still have meaning. The Environmental Data Registry documents data elements. The purpose of documenting data elements is to encourage the standardization of data among projects. The EDR is consistent with ISO 11179, an international standard that provides guidance on the formulation and maintenance of discrete data element descriptions and metadata that can be used to describe data elements in a consistent manner. The EDR is currently under development and is expanding rapidly.

When fully developed, the EDR is intended to be a comprehensive, authoritative source of reference information about environmental data. It is not the environmental data itself, but the information that helps describe the data and make it more meaningful. The EDR serves as the clearinghouse for information about the data. It provides information on the definition, origin, source, and location of environmental data. When used in conjunction with an environmental information database, the EDR enables users to better understand the information they are accessing. It also serves as a major tool to support a standard-setting process, to record and disseminate these standards and to facilitate data sharing between organizations and users.

The Environmental Data Registry provides a single source of information about EPA data, fulfilling the functions of dictionary, directory, repository and standardization tool. The EDR functions as a dictionary by providing information on names, definitions, and formats of environmental data. It serves as a directory by identifying responsible parties and contact information for data. The EDR is a repository of required data elements for database system design, documenting the history of data, including its accuracy and precision. It provides location and relationship information such as the applications in which the data is used and databases in which data is maintained, thereby serving as a tool for data standardization.

### *2.1.4 Database Documentation*

A database may be viewed as a collection of information that is organized so that its contents can be easily accessed, managed, and updated. Data modeling organizes data into a structure that represents the real world as closely as possible and allows the information to be processed by computers. A data model is a conceptual representation of the data structures that are required by a database. The data structures include the data objects, the associations between data objects, and the rules that govern operations on the objects. The data model is equivalent to an architect's building plans.

Data models are documented through schematic diagrams, such as Entity-Relationship (E-R) diagrams, that pictorially represent the entities which comprise the database and the relationships between them. An entity is a ‘thing’ which can be distinctly identified, for example a person, a car, or an event. A relationship is an association among entities; e.g. “person OWNS car” is an association between a person and a car. By representing entities, their cardinality relationships (e.g., one-to-one, one-to-many, etc.), and the keys that connect them, an E-R diagram is both a design tool and a map that shows how to use the data.

Data models are translated into a form that can be used by a computer through incorporation into a database. Other users will need to understand the structure of the database as well as the entities it contains. While the structure is described by the E-R diagram described above, entities are documented through the development of a data dictionary.

A data dictionary is a collection of descriptions of the entities or items in a data model for the benefit of programmers and others who might need to refer to them. After each data object or item is given a descriptive name, its relationship is described, the type of data (such as text or image or binary value) is described, possible predefined values are listed, and a brief textual description is provided. This collection becomes the data dictionary. When developing programs that use the data model, a data dictionary can be consulted to understand where a data item fits into the structure, what values it may contain, and what the data item means in real-world terms.

## 2.2 Documentation Requirements for EMPACT Projects

Project documentation for EMPACT projects is intended to facilitate the sharing and understanding of information. Documentation at the project level will allow other communities to benefit from the experience of EMPACT projects and develop their own projects. Documentation of data sets will assist secondary users of that data in understanding the data. For example, someone wanting to expand an ozone map to a larger region will be able to determine whether different sets of data can be compared and projected onto the same map. People who travel from one location to another can determine whether data from these two locations have the same meaning. Table 2-1 lists the types of documentation that are considered relevant to EMPACT projects.

**TABLE 2-1:** Project Documentation

| Documentation Relevant to EMPACT Projects | |
|---|---|
| **Project Level** | Report describing what the project is trying to accomplish, the process being used, and the resources required to undertake the project |
| **Data Set Level** | • **Metadata providing what, where, when, who and how of data set**<br>• **Raw data files** |
| **Data Element Level** | Use Environmental Data Registry to select elements from which to build the database and record new elements in the EDR |
| **Database** | Data Dictionary |

Failure to adequately document a project will reduce its usefulness. Other users will not have information they need to duplicate successful projects and will have to start from scratch in developing their own projects. There will be no basis for comparison between different sources of data that will serve as a barrier to aggregation of data.

### 2.2.1 Project Level Documentation

For EMPACT projects, documentation at the project level would consist of a text report that includes the following information:

- ✔ Project Title
- ✔ Purpose
- ✔ Project Description
- ✔ Methodology Used
- ✔ Point of Contact
- ✔ Funding

This report should be provided to the EMPACT project coordinator at the conclusion of the project development phase. It should also be made available from the project Web site.

### 2.2.2 Data Set Level Documentation

The following list constitutes a minimal adequate set of metadata needed to document data sets for general "catalog" use. This information should be provided for each data set generated in EMPACT projects and made available on the project Web site. Assistance is available to EMPACT teams for creating this metadata. This assistance consists of a Web form for facilitating creation of this information and is described in Section 2.3.

***Repeating "Profile" Elements***

The following elements are collected once for a given metadata entry person.

**General Contact Information**

**Organization:**

**Metadata contact position/person:**

**Address type:**

**Address:**

**City:**

**State or province:**

**Postal code:**

**Country:**

**Phone:**

**Fax:**

**E-mail address:**

***Metadata Elements for Each Entry***

The following elements are generated for each data set created for the EMPACT project and made available on the project Web site.

## General Identification Information

**Identity of this entry** (for future update):

**Originator:** the name of an organization or individual that developed the data set

**Publication date** (YYYYMMDD): the date when the data set is made available for release

**Title of data set:** the name by which the data set is known

**Edition:** the version of the title

**Publication place:** the name of the city (and state or province, and country, if needed to identify the city) where the data set was published or released

**Publisher:** the name of the individual or organization that published the data set

**Online linkage (URL):** the URL where the data may be found

**Abstract:** a brief narrative summary of the data set

**Purpose:** a summary of the intentions with which the data set was developed

**Supplemental Information:** other descriptive information about the data set

**Beginning date:** (YYYYMMDD): the first year, month, and day of the event

**Ending date:** (YYYYMMDD): the last year, month, and day of the event

**Progress:** the state of the data set

**Intended data set maintenance and update frequency:** the frequency with which changes and additions are made to the data set after the initial data set is completed

**West bounding coordinate (-DDD.XXX):**

**East bounding coordinate (-DDD.XXX):**

**North bounding coordinate (DD.XXX):**

**South bounding coordinate (DD.XXX):**

**Theme keywords:** common-use word or phrase used to describe the subject of the data set

**Place keywords:** the geographic name of a location covered by a data set

**Limits on data accessibility:** restrictions and legal prerequisites for accessing the data set

**General Identification Information**

**Limits on use of data:** restrictions and legal prerequisites for using the data set after access is granted

**Distribution Information**

**Distribution organization:**

**Distribution contact position/person:**

**Address type:**

**Address:**

**City:**

**State or province:**

**Postal code:**

**Country:**

**Phone:**

**Fax:**

**E-mail:**

**Data set name as known by Distributor:**

**Liability held by distributor:**

**Date of last metadata entry or update (YYYYMMDD):**

***Sample Metadata***

An example of a set of fictitious metadata can be found in the National States  Geographic Information Council Metadata Primer at *http://www.lic.wisc.edu/metadata/acmemin.htm*.

### 2.2.3 Use of the Environmental Data Registry on EMPACT Projects

One of the main goals of using the EDR in conjunction with EMPACT projects is to standardize the use of data elements to ensure the comparability of data across projects.  The common elements of all EPA reporting requirements will be translated into data standards registered in the EDR. The EDR will identify reusable data within and between organizations and enable access and understanding of environmental data by users. It will be the source of documentation about data to aid in human understanding, preserve meaning over time, and facilitate intelligent software processing.

Use of the EDR on EMPACT projects will enable the sharing of information between projects by serving as a tool for standardization of data elements. EMPACT project developers who are beginning development of their databases should begin their efforts by searching the EDR for elements relevant to their project. These elements can serve as the basis for database creation.

**Steps in using the EDR on EMPACT projects:**

1. Begin database development by going to the EDR and finding elements that can be used in constructing the database
2. Register any elements used in the database which do not currently exist in the EDR

Elements recommended for use in EMPACT projects have been documented in the *Summary Report of Proposed Standard Data Elements for EMPACT Projects* prepared by the EPA Systems Development Center. A link to this document is available on the EMPACT Web site. Any elements needed that can not be located in the EDR would then require registration. Assistance in finding good elements can be found on the EDR Web site at URL *http://www.epa.gov:6706/edrdcd/owa/helptip$help.QueryView?p_help_id=62*. EMPACT project developers who already have databases should search the EDR to determine if there are any data elements used in their projects which are not currently existing in the EDR. Such elements should be registered. Searches can be performed from the EDR Web site at URL *http://www.epa.gov:6706/edrdcd/owa/BROWQRY$.STARTUP*.

### *2.2.4 Database Documentation*

EMPACT projects should document their database development so that the data, as well as the database structures, can be utilized by other users. A data dictionary, described in Section 2.1.4 above, that describes the entities used in the database, should be developed and made available on the project Web site.

## 2.3 EPA Assistance Available to EMPACT Projects

A form for generating metadata at the data set level will be available on the EMPACT Web site (http://www.epa.gov/empact/). Completion of this form will generate an html page listing the metadata that will then be emailed back to the recipient for inclusion on their Web site.

Assistance is also available for finding or registering elements in the Environmental Data Registry. Users comfortable with the process may submit electronic files or complete a web-based form. Both methods are described on the EDR Web site at *http://www.epa.gov:6706/edrdcd/owa/REG$.STARTUP*. Further assistance in finding and defining elements and registering them in the EDR can be obtained by contacting the EPA Systems Development Center (703-908-2474).